DOCUMENT RESUME

ED 478 457 CS 512 074

AUTHOR	Lubliner, Shira; Smetana, Linda
TITLE	Recognition or Recall: What Reading Comprehension Tests Really Measure.
PUB DATE	2003-04-22
NOTE	11p.; Paper presented at the Annual Meeting of the American Educational Research Association (84th, Chicago, IL, April 21-25, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Academic Achievement; *Constructed Response; Gender Issues; Grade 5; Intermediate Grades; *Multiple Choice Tests; Program Evaluation; *Reading Comprehension; Reading Tests; *Validity

ABSTRACT

This study examined format differences in the measurement of fifth grade students' reading comprehension achievement. Children were given a multiple-choice reading comprehension test, followed 4 weeks later by a constructed response test on 2 of the same text passages. Results indicated that little comprehension of text content was transferred from the multiple-choice to the constructed response test. Significant gender differences were detected in overall test performance and consistency across test formats. Results raise concerns regarding the validity of multiple-choice tests in the measurement of children's reading comprehension proficiency. (Contains 26 references, 4 tables, and 2 figures.) (Author/RS)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 478 457

Recognition or Recall:

What Reading Comprehension Tests Really Measure

Annual Meeting of the American Educational Research Association

April 22, 2003

Shira Lubliner

and

Linda Smetana

California State University, Hayward

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

S. Lubliner

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

CS 512 074

BEST COPY AVAILABLE

Abstract

This study examined format differences in the measurement of fifth grade students' reading comprehension achievement. Children were given a multiple-choice reading comprehension test, followed four weeks later by a constructed response test on two of the same text passages. Results indicated that little comprehension of text content was transferred from the MC to the CR test. Significant gender differences were detected in overall test performance and consistency across test formats. Results raise concerns regarding the validity of MC tests in the measurement of children's reading comprehension proficiency.

Purpose

Reading achievement has been increasingly politicized in recent years; resulting in intense pressure on schools to improve standardized test scores. High stakes tests are used for accountability purposes, based on the assumption that scores reflect meaningful reading competencies. Children who select a high percentage of correct answers on multiple-choice (MC) reading comprehension tests are judged to be proficient in reading comprehension. MC reading tests are widely used due to their purported reliability and cost-effectiveness. But questions have been raised regarding the validity of these tests in measuring the reading comprehension skills that children use in natural reading activities. To what extent do MC and constructed response (CR) tests measure the same underlying construct, reading comprehension? A number of studies have examined the relationship between test format and student performance with mixed results (Mehrens, 1991). Substantial differences between MC and CR test scores have been documented in some studies (Shavelson et al., 1992; Shepard, 1997). Other studies have provided evidence of near perfect correlations between test formats after scores have been corrected for attenuation (Bennett et al., 1991; Bridgeman & Lewis, 1994; Bridgeman & Rock, 1993; Lukhele et al., 1994; Traub, 1993). None of these studies focused on elementary school children's reading comprehension achievement.

Numerous researchers have pointed out the problems associated with the MC format. Messick (1989, 1992) found that children's varying levels of test-taking skills make MC tests susceptible to construct-irrelevant difficulty or easiness, threatening test validity. Shepard (1997) suggested that the distortion of MC scores due to test-wiseness, allows some students to demonstrate knowledge they don't actually have. According to Kintsch (1998) the question-answer format does not tap deep-level understanding of texts. He recommended against the use of questions in scientific studies of comprehension due to invalid inferences that may result from such tests. In addition to concerns about validity, a number of studies have found that MC tests provide an advantage to male students (Breland et al., 1994; Lane, Wang, & Magone, 1995; Poplun & Capps, 1999).

Educators agree that reading comprehension proficiency must be measured accurately if educators are to evaluate student progress, determine effective methods of literacy instruction, and hold schools accountable for student learning. The psychometric quality of standardized MC tests is cited as a justification for using these tests for high-stakes purposes. A number of states

such as California rely on MC tests as a single measure of reading achievement, providing schools with rewards and sanctions based entirely on students' performance on the SAT 9 test. The over-reliance on standardized MC test scores highlights the importance of investigating this test format. The purpose of this study was to examine format differences in the measurement of fifth grade students' reading comprehension to determine whether MC tests accurately measure reading comprehension skills. Additionally, the study was designed to explore the association between gender and test format to determine whether MC reading comprehension tests provide an advantage to one gender over another.

Theoretical Rationale

Cognitive theorists, Kintsch and Van Dijk (1978) developed a model for conceptualizing the procedural abilities that underlie reading comprehension. They theorized that comprehension involves a complex series of activities including the organization of text into a coherent whole, the condensation of meaning into a gist and generation of new texts, summaries, based on recall and reconstruction. The process of comprehension is dependent on the identification of key discourse structures known as propositions. According to Kintsch's construction-integration (CI) model, text representations are developed sequentially in a word-by-word, sentence-by-sentence process as meaning is constructed (1998). The conceptualization of reading as a constructive, meaning-making process is widely accepted among reading researchers (Brown, 1985; Cross & Paris, 1988; Palincsar, 1985; Pearson & Dole, 1987; Pressley et al., 1989; Ruddell & Unrau, 1994; Rumelhart, 1980, 1981). The CI model provides a framework for understanding how reading comprehension works and how it is measured by MC and CR tests.

Method

Participants and Setting

The study took place in an affluent suburban elementary school located in Northern California. Seventy-two percent of the students are white, 16% are Asian, 5% are Latino. The students' 2001 SAT 9 reading scores averaged in the 80th percentile. One hundred and twenty-five students in four intact fifth grade classes participated in the study. Participants included 62 boys and 63 girls.

Research Design

The research design was correlational, designed to explore the association between two sets of reading comprehension tests. The MC tests, developed for a prior study, consisted of four 200-word social studies passages and corresponding 10-item MC item sets. A total of 30 items were designed to measure reading comprehension and 10 items measured vocabulary achievement. The CR tests consisted of two of the same 200-word social studies passages that were found on the MC test that the students were required to read and summarize. Scoring of the CR tests was based on 10 important propositions that the researchers identified in each of the two text passages.

The researchers conducted all assessment activities in the students' regular classrooms. The MC test was administered first. Four weeks later the researchers returned to the classrooms and administered the CR tests. Students were given the first passage and asked to read it. The

passages were collected and students were asked to write a detailed summary (narrative or bullet-point format was accepted.) The process was repeated with the second passage.
The researchers rated the summaries based on the list of propositions that had been generated. Forty tests were randomly selected for calibration of scoring. The researchers scored the summaries, compared scores and resolved disagreements by discussion. Inter-rater reliability averaged .85 on the scoring of the two summaries.

Results

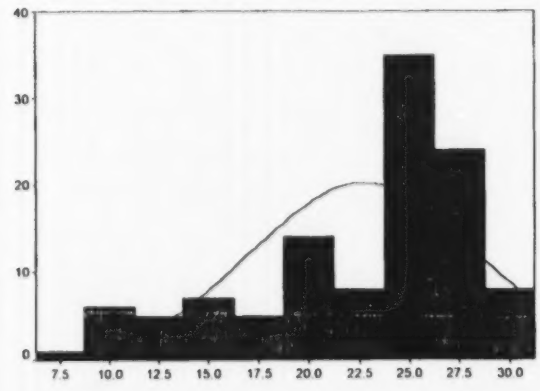
Results of the analysis indicated that the MC mean was 22.72 (30 items on the test) and the standard deviation was 5.57. The CR mean was 9.27 (20 propositions on the test) and the standard deviation was 3.21. Scores were converted into percentages to allow for a meaningful comparison across test formats. Results indicated that student achievement was much higher on the MC test (76%) than on the CR test (46%).

Table 1
MC and CR Scores

	N	*Mean*	*Std. Deviation*	*Percent*
MC	113	22.72	5.57	.76
CR	113	9.27	3.20	.46

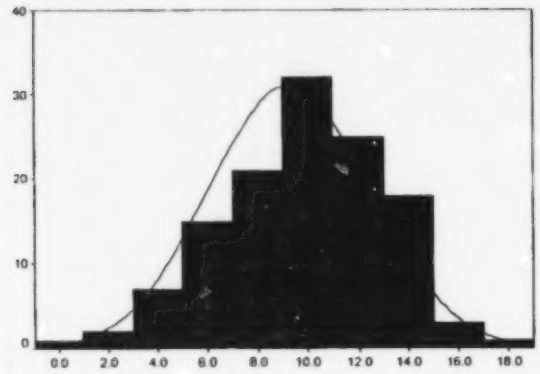
An analysis of histograms for the two tests was particularly interesting. The MC test scores produced highly skewed results (-.98, std error .23), while the CR test produced a slightly skewed pattern more closely resembling normal curve (-.32, std error .22). The histogram for the MC test suggests that the test items may have been too easy for many of the students in this high-performing school, creating a ceiling effect. This pattern is not evident on the CR histogram. The lack of symmetry in the histograms suggests a format-related advantage for high-performing students associated with the MC test.

Figure 1
Histogram for MC Test Scores



BEST COPY AVAILABLE

Figure 2
Histogram for CR Test Scores



CR Comprehension

When data were disaggregated by gender, results indicated that girls outperformed boys on all measures, with significant achievement differences on the CR test.

Table 2
MC and CR Test Scores Disaggregated by Gender

		MC Test		*CR Test*	
	N	*Mean*	*Std. Deviation*	*Mean*	*Std. Deviation*
Boys	52	21.87	5.71	8.58	3.14
Girls	61	23.44	5.40	9.95	3.15

Gender differences were also substantial when correlations between MC and CR tests were examined. The correlation between test scores was .48 for boys and .74 for girls (significant at the .01 level.)

Table 3
Correlation Between MC and CR Scores

	N	
Boys	52	.48**
Girls	61	.74**
Total	113	.57**

**Significant at the .01 level

A regression analysis was performed with the MC reading comprehension and vocabulary scores used as predictors of CR test achievement. R_2 was calculated for each gender, indicating that the

BEST COPY AVAILABLE

predictors explained only .26 of the variance of boys' scores compared to .55 of the variance in girls' scores. Results indicated that girls' performance was more consistent across test formats.

Table 4

Regression Analysis

	R	*R2*	*Adjusted R2*	*Std. Error*
Boys	.51	.26	.23	2.82
Girls	.74	.55	.53	2.19
Total	.64	.41	.39	2.52

Dependent variable: CR
Predictors: MC comprehension & MC vocabulary

Discussion

Results of this study suggest that the process of generating propositions and constructing summaries on the CR test was very difficult for the students despite the fact that they had been tested on both text passages four weeks earlier. The MC format may have been easier because it provided continuous access to the text as children selected answers to test items. Although format differences may account for the high MC scores, it is difficult to ascertain why exposure to the texts on the MC test did not inflate CR scores. The fact that overall scores were much lower on the CR test suggests that children transferred very little information from the MC test. Kintsch (1998) explains that MC items measure children's ability to recognize, rather than recall information and that recognition does not require the deep level processing necessary for the construction of a summary. Results of this study suggest that the MC test measured recognition rather than text recall or reading comprehension.

When children completed items on MC and CR tests distinct differences in test taking behavior were observed. When given the MC test many children skipped reading the text altogether and went directly to the test items. They read the question and skimmed over the text until a response was located, repeating the process for each of the 30 comprehension items. The "skim and recognize" activities observed in the MC test may have been effective in producing high test scores but reflected little engagement with the text and few of the behaviors have been identified as integral to the reading comprehension process.

The test-taking behaviors observed on the CR test were quite different than those observed on the MC test. When given the text for the CR test nearly all of the children appeared engaged in reading. When they began writing their summaries, children listed propositions in the exact order that they occurred in the text. The children's performance on the CR test supported Kintsch's theory that texts are recalled through a series of propositions and that comprehension is constructed in a sequential manner.

Girls' reading comprehension achievement was substantially more consistent than that of boys across test formats. According to Shepard (1997) variability in test performance results from children's fragile understanding and inability to generalize knowledge from one format to another. To follow Shepard's reasoning, it is possible that the girls in this study had learned to

see beyond the differences in test format, discerning the deeper structure in reading comprehension tasks. The less consistent performance of boys may have resulted from greater reliance on a "skim and recognize" strategy for identifying the correct answers on MC test. In short, boys' MC scores may reflect test-taking skills rather than reading comprehension proficiency.

Educational Importance

These results suggest the importance of teaching children the cognitive processes that underlie reading comprehension. When children learn how to generate propositions, formulate a gist of a text, fill in inconsistencies with inference, and construct a summary of what has been read they are developing authentic reading comprehension skills that transcend test format (Kintsch, 1998). Children become proficient at reading comprehension when they are taught to construct understanding that can be transferred from one context to another (Tierney & Pearson, 1994). The emphasis on test-taking skills, on the other hand, teaches children to look for correct answers rather than understanding. As Shepard (1997) points out, an emphasis on MC test preparation may reduce children's ability to generalize knowledge in addition to reinforcing ineffective reading habits.

These results highlight a serious deficiency in the way children's reading comprehension achievement is measured. Most high stakes decisions regarding school accountability are based on standardized MC tests. These tests may not be a valid measure of reading comprehension achievement, particularly for boys. MC tests appear to inflate boys' scores substantially more than those of girls, a problem that is accentuated, as students grow older. It is likely that MC tests are a flawed method of measuring children's reading comprehension proficiency and should not be used for high stakes accountability purposes.

References

Bennett, R., Rock, D., & Wang, M. (1991). Equivalence of free-response and multiple- choice items. *Journal of Educational Measurement, 28*, 77-91.

Breland, H., et al. (1994). Performance versus objective testing and gender: an exploratory study of an advanced placement history examination. *Journal of Educational Measurement, 31.*

Bridgeman, B. & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement,31*, 37-50.

Bridgeman, B. & Rock, D. (1993). Relationship among multiple-choice and open-ended analytical questions. *Journal of Educational Measurement, 30*, 313-329.

Brown, A. (1985). *Teaching students to think as they read: Implications for Curriculum reform.* American Educational Research Association Task Force on Excellence in Education, (HD-06864): Curriculum Reform.

Cross, D. & Paris, S. (1988). Developmental and instructional analysis of children's metacognition and reading comprehension. *Journal of Educational Psychology 80* (2), 131-142.

Haladyna, T. (April, 1998). *Fidelity and proximity to criterion: When should we use multiple-choice?* Address presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Kintsch, W. (1998). *Comprehension, A paradigm for cognition.* Cambridge, UK: Cambridge University Press.

Kintsch, W. & van Djk, T. (1978). Toward a model of text comprehension and production. *Psychological Review, 85* (5), 363-394.

Lane, S. Wang, N. & Magone, M. (April,1995). *Gender-related differential item functioning on a middle-school mathematics performance assessment.* Address presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement, 31*, 234-250.

Mehrens, W. (1991). *Using performance assessment for accountability purposes: Some problems.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, Il.

Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational measurement* (3rd ed., p. 13-104).

Messick, S. (1992). *The interplay of evidence and consequences in the validation of performance assessments*. Invited address to the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Palincsar, A. (1985). *The unpacking of a multi-component, metacognitive training package.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Il.

Pearson, D. & Dole, J. (1987). Explicit comprehension instruction: A review of research and a new conceptualization of instruction. *Elementary School Journal, 88* (2)151-165.

Poplun, M. & Capps, L. (1999). Gender differences for constructed-response mathematics items. *Educational and Psychological Measurement, 59*, 597-614.

Pressley, M., Johnson, C., Symons, S., McGoldrick, J., Kurita, J. (1989). Srategies that improve children's memory and comprehension of text. *The Elementary School Journal 90* (1), 3-32.

Ruddell, R. & Unrau, N. (1994). Reading as a meaning-construction process: The reader, the text and the teacher. In R. Ruddell, M. Ruddell, & H. Singer (Eds.) *Theoretical models and processes of reading*, 4th edition. Neward, DE: International Reading Association.

Rumelhart, D. (1980). *An introduction to human information processing*. New York: Wiley.

Rumelhart, D. (1981). Schemata: The building blocks of cognition. In J.T. Guthrie (Ed.) *Comprehension and teaching: Research reviews* (p. 3-26). Newark, DE: International Reading Association.

Rumelhart, D. (1994). Toward an interactive model of reading. In R. Ruddell, M. Ruddell, & H. Singer (Eds.) *Theoretical models and processes of reading*, 4th edition. Neward, DE: International Reading Association.

Shavelson, R., Baxter, G. & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher*, p. 22-27.

Shepard, L. (1997). *Measuring Achievement: What does it mean to test for robust understanding?* Paper presented at the annual William H. Angoff Memorial Lecture, Educational Testing Service, Princeton, NJ.

Tierney, R. & Pearson, D. (1994). Learning to learn from text: A framework for Improving classroom practice. In R. Ruddell, M. Ruddell, & H. Singer (Eds.) *Theoretical models and processes of reading*, 4th edition. Neward, DE: International Reading Association.

Traub, R. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In Bennett, R. and Ward, W. (Eds.), *Construction Versus Choice in*

Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum Associates.

CS 512 074

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Recognition or Recall: What Reading Comprehension Tests Really Measure

Author(s): Shira Lubliner Ed.D. and Linda Smetana Ed.D.

Corporate Source: Presentation at the Annual Meeting of the American Education Research Association

Publication Date: Presentation Date April 22, 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: Shira Lubliner

Printed Name/Position/Title: Shira Lubliner Ed.D.

Organization/Address: California State University Hayward
25800 Carlos Bee Blvd
Hayward, CA 94542

Telephone: (510) 885-4414

FAX:

E-Mail Address: slubline@csuhayward.edu

Date: 6/4/03

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distnbutor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org

EFF-088 (Rev. 2/2001)